Untitled

"Saba_Alemayehu"

6/20/2022

Exploratory Data Analysis

```
stroke_dt<- read.csv("healthcare-dataset-stroke-data.csv")
class(stroke_dt)
```

```
## [1] "data.frame"
```

```
colnames(stroke_dt)
```

```
##  [1] "id"                "gender"            "age"
##  [4] "hypertension"      "heart_disease"     "ever_married"
##  [7] "work_type"         "Residence_type"    "avg_glucose_level"
## [10] "bmi"               "smoking_status"    "stroke"
```

```
str(stroke_dt)
```

```
## 'data.frame':    5110 obs. of  12 variables:
##  $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
##  $ gender           : chr  "Male" "Female" "Male" "Female" ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(stroke_dt)
```

```
##        id           gender               age         hypertension
##  Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Mode  :character   Median :45.00   Median :0.00000
##  Mean   :36518                      Mean   :43.23   Mean   :0.09746
```

```
##  3rd Qu.:54682                     3rd Qu.:61.00   3rd Qu.:0.00000
##  Max.   :72940                     Max.   :82.00   Max.   :1.00000
##  heart_disease     ever_married        work_type         Residence_type
##  Min.   :0.00000   Length:5110        Length:5110        Length:5110
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.05401
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  avg_glucose_level     bmi             smoking_status         stroke
##  Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000
##  1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000
##  Median : 91.89    Mode  :character   Mode  :character   Median :0.00000
##  Mean   :106.15                                          Mean   :0.04873
##  3rd Qu.:114.09                                          3rd Qu.:0.00000
##  Max.   :271.74                                          Max.   :1.00000
```

```
head(stroke_dt)
```

```
##      id gender age hypertension heart_disease ever_married     work_type
## 1  9046   Male  67            0             1          Yes       Private
## 2 51676 Female  61            0             0          Yes Self-employed
## 3 31112   Male  80            0             1          Yes       Private
## 4 60182 Female  49            0             0          Yes       Private
## 5  1665 Female  79            1             0          Yes Self-employed
## 6 56669   Male  81            0             0          Yes       Private
##   Residence_type avg_glucose_level  bmi  smoking_status stroke
## 1          Urban            228.69 36.6 formerly smoked      1
## 2          Rural            202.21  N/A    never smoked      1
## 3          Rural            105.92 32.5    never smoked      1
## 4          Urban            171.23 34.4          smokes      1
## 5          Rural            174.12   24    never smoked      1
## 6          Urban            186.21   29 formerly smoked      1
```

```
#tail(stroke_dt)
#apply(stroke_dt,2,class)
```

encoding the variables into numeric and factors

```
stroke_dt$id<-as.numeric(stroke_dt$id)
stroke_dt$age<-as.numeric(stroke_dt$age)
stroke_dt$bmi <- as.numeric(stroke_dt$bmi)
```

```
## Warning: NAs introduced by coercion
```

```
stroke_dt$avg_glucose_level<-as.numeric(stroke_dt$avg_glucose_level)


stroke_dt$hypertension<-as.factor(stroke_dt$hypertension)
```

```
stroke_dt$heart_disease<-as.factor(stroke_dt$heart_disease)

stroke_dt$gender <- as.factor(stroke_dt$gender)
stroke_dt$work_type<-as.factor(stroke_dt$work_type)
stroke_dt$ever_married <- as.factor(stroke_dt$ever_married)
stroke_dt$Residence_type <- as.factor(stroke_dt$Residence_type)
stroke_dt$smoking_status <- as.factor(stroke_dt$smoking_status)
stroke_dt$stroke <- as.factor(stroke_dt$stroke)

class(stroke_dt$stroke)
```

```
## [1] "factor"
```

```
# change the level name of response variable as "",""
levels(stroke_dt$stroke)=c("no","yes")
```

check the missing values

```
apply(stroke_dt,2,function(x)sum(is.na(x)))
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0               201                 0                 0
```

```
table(stroke_dt$gender)
```

```
##
## Female   Male  Other
##   2994   2115      1
```

```
table(stroke_dt$hypertension)
```

```
##
##    0    1
## 4612  498
```

```
table(stroke_dt$heart_disease)
```

```
##
##    0    1
## 4834  276
```

```
table(stroke_dt$ever_married)
```

```
##
##   No  Yes
## 1757 3353
```

```
table(stroke_dt$Residence_type)
```

```
##
## Rural Urban
##  2514  2596
```

```
table(stroke_dt$smoking_status)
```

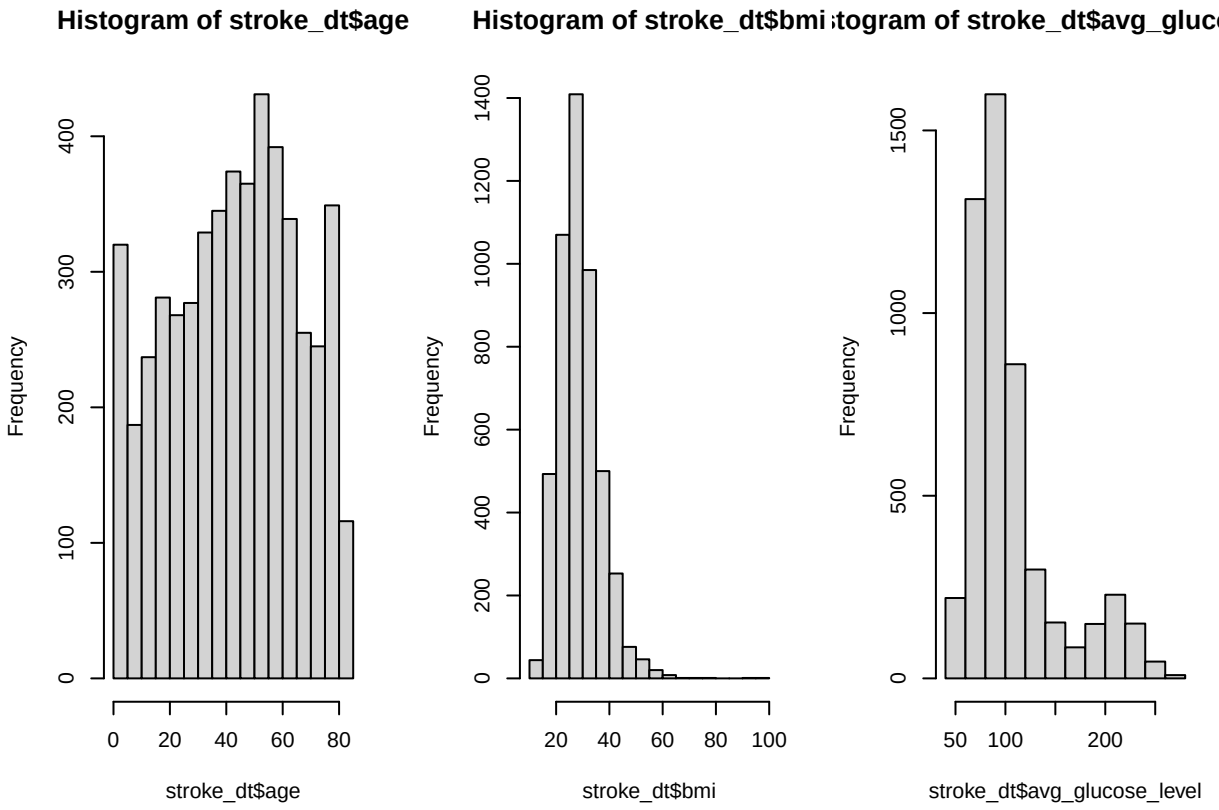
```
##
## formerly smoked    never smoked          smokes         Unknown
##             885            1892             789            1544
```

```
table(stroke_dt$stroke)
```

```
##
##   no  yes
## 4861  249
```

Data Visualization

```
# histogram of numeric variables
par(mfrow=c(1,3))
hist( stroke_dt$age )
hist( stroke_dt$bmi )
hist( stroke_dt$avg_glucose_level )
```



Class Imbalance

```
barplot(prop.table(table(stroke_dt$stroke)),
        col = rainbow(2),
        ylim = c(0, 1),
        main = "Class Distribution")
```

Data Spliting and data preprocessing

```
#first separate the response and predictor variables
strok_predictor <- subset(stroke_dt, select = -c(stroke))
stroke_response<-stroke_dt$stroke
```

#Data partitioning

```
set.seed(100)
training_rowSt<-createDataPartition(stroke_response, p=0.8, list=FALSE)

strok_prTrainX<-strok_predictor[training_rowSt,]
strok_prTestX<-strok_predictor[-training_rowSt,]
#str(strok_prTrainX)

stroke_reTrY<-stroke_response[training_rowSt]
stroke_reTesY<-stroke_response[-training_rowSt]
#str(stroke_reTrY)
```

Data Preprrocessing

```
# Check for zero variance predictors
StZero_coln<- nearZeroVar(strok_prTrainX)
str(StZero_coln)
```

```
##  int(0)
```

```
# there is no variables which have near zero or zero variance
```

Impute the missing value

```
trainimpu<-preProcess(strok_prTrainX,"knnImpute")
strokeTrprX<-predict(trainimpu,strok_prTrainX)
#str(strokeTrpr)
strokeTeprX<-predict(trainimpu,strok_prTestX)
```

Develop a model

```
library(caret)
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.1.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
#Logistic Regression
ctrl<-trainControl(method="LGOCV",
                   summaryFunction=twoClassSummary,
                   classProbs = TRUE,
                   savePredictions = TRUE,
                   sampling = "up")


set.seed(300)
#lrSfit<-train(x=strokeTrprX, y=stroke_reTrY,
              #method = "glm",
              #metric="ROC",
```

```
              #preProcess = c("center","scale"),
              #trControl=ctrl)

#lrSfit
```

```
#MDA
```

```
#Neural Networks
```

```
#Support Vector Machines
```

```
#K-Nearest Neighbors
```

```
#Naive Bayes
```

```
#Random Forests
```

```
#Boosting
```